International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-5, January 2019

Intelligent Heart Disease Prediction using Neural Network

Soumonos Mukherjee, Anshul Sharma

Abstract*: Health-care is a field of the most needed service and an economically 2nd largest industry in 21st century. While we talk about the affordability and quality assurance in health-care industry, several statistical analysis is carried on to make health solutions more precise and flawless in this current era of increasing health problems and chronic diseases. Advancements on data driven intelligent technologies is disease diagnosis and detection, treatment and research are remarkable. Medical image analysis, symptom based disease prediction is the part where the most sought after brains are working. In this paper we aim to present our proposed model on the prediction on diagnosis of cardio vascular disease with ECG analysis and symptom based detection. The model aims to be researched and advance in further to become robust and end to end reliable research tool. We will discuss about the classical methods and algorithms implemented on CVD prediction, gradual advancements, draw comparison of performance among the existing systems and propose an enhanced multi-module system performing better in terms of accuracy and feasibility. Implementation , training and testing of the modules have been done on datasets obtained from UCI and Physionet data repositories. Data format have been modified in case of the ECG report data for betterment of action by the convolutional neural network used in our research and in the risk prediction module we have chosen attributes for training and implementing the multi-layered neural network developed by us. The further research and advancement possibilities are also mentioned in the paper.

***Key words:** attribute, classification CVD(Cardio vascular disease), convolutional neural network, multi-layered neural network, Physionet, UCI (University of California,Irwin).*

I. INTRODUCTION

Data mining is the field of storing, structuring and analysing massive scale historical and authenticated data to find the frequent or very unexpected patterns and correlations among the part objects which seem to be unrelated and iteratively continue the paradigm style approach to get the knowledge derived from the huge amount of data on a scalable manner. As the disease diagnosis is known to be the most crucial part of clinical medicine, application of data driven methods and using the doctors' knowledge and experiences to design the machine learning algorithms to carry over the patients' data makes it easier, time efficient, affordable and more accurate. So hereby accuracy of the prediction is the prime concern of ours while going through a predictive methodology and model.

II. LITERATURE SURVEY:

We have cite d and reviewed many of the existing systems proposed in the previously published research articles and drew a comparative analysis based upon the performance matrices produced by their approaches. Here we mentioned some significant ones.Yadav,Tomar and Agarwal in their paper [1] implemented the foggy K means clustering for predicting disease pattern in a large real time dataset (which includes missing values and high dimensional data and proved that an alternate classification method like foggy k means comes out giving a really good prediction accuracy. Nazeer, Naveed and Akram [2] used homogenous ensemble of SVMs for the feature classification of disease dataset and the final result is optimized and checked on accuracy by genetic algorithm. Improves the result reliability quite more than the existing systems. Jyothi et al [3] generated a good number of association rule algorithms to find the frequent pattern in the dataset. Algorithms like A-priory and FP-growth has been frequently used in their research. Ade et al proposed [4] the disease prediction using SVM and naive bayes classifier, make flags of very high, high, low, very low classifications and any unknown constraint would fit in based on the approach.Ibrahim et al used [5] three term backpropagation network model and tuned to accuracy with one of the multiobject genetic algorithm(MOGA) i.e: nondominated sorting genetic algorithm which simultaneously also increases the reliability of the classification. Many of the popular prediction methods [6] i.e: logistic regression, SVM and random forest models suffer from the problem of singular prediction rule and lack the capacity of integrating different rules generated on the basis of differential interventions of the patients. Anbarasi et al implemented [7] and applied the genetic algorithm for improving the accuracy by classifying and defining subset features in disease database. Nidhi bhatla et al [8] used certain data mining methods and showed that decision tree is a good method of production when associated with genetic algorithms and considered on 9 and 14 attributes respectively.Some of the studied models are illustrated below:

K means clustering: The one is one of the most extensively used data mining algorithmic model used in the island-type segregation and group of similar data (which is called clustering) from a chunk of unlabelled dataset following an unsupervised learning principle.

Revised Manuscript Received on 30 January 2019.
***** Correspondence Author
Soumonos Mukherjee, Department of Computer Science and Engineering VIT,Vellore, Tamil Nadu India.
Anshul Sharma, Department of Computer Science and Engineering VIT,Vellore, Tamil Nadu India.

This is widely used in finding the probability of heart disease from a large enough patient dataset with the given input attributes (e.g: Headache, chest pain, Blood pressure level etc.).

A-priory: This algorithm is the first developed algorithm of data science and has an wide range of usage. There are several algorithms derived from the nascent A-priory algorithm and they are being effectively used for determining frequent patterns in large unstructered datasets.Some of the early stage disease prediction papers were written on methodologies using A-priory and FP growth on patient datasets. It showed quite improved result on prediction.

Decision tree: A very widely used machine learning predictive tool is decision support tool or decision tree which particularly a tree type acyclic graph structure to analyse particular dataset by pruning the given attributes on a likelihood measure to find the measurement of prediction which are appended as leaf nodes. There are several developments on various categories of this type such as ID3, C4.5 etc. The former one uses entropy minimization and information gain maximization paradigm finding the result with the minimal variance (towards maximum accuracy). In healthcare analytics and predictive clinical diagnosis decision tree based techniques are extensively used. Support vector machines: Support vector machines are unsupervised classification techniques working by first projecting data points on a hyper plane and then dividing it into two classes maximizing the distance between the hyper plane and two closest data points from each of the class and thus minimizing the error relating misclassification. This is used for heart disease predictions associated with the naive bayes classifier for considerable improvement of prediction accuracy than the primer techniques. Naive Bayes: A probabilistic classifier building method which uses Bayes's theorem of conditional probability: P(X|C) = P(C|X) * P(C) / P(X) -for classification. X signifies the data tuples and C indicates the class. To be considered that X is uniform for all C and moreover all the attribute values are conditionally mutually inter dependent. It performs quite good along with SVM to predict patterns on large dataset in a time efficient manner.

Genetic Algorithm: Genetic algorithm is the brand new invention for neural network based artificial intelligent systems. Genetic algorithms simulate darwininan theory of evolution based on natural selection. It materializes the characters and propagation of inherent typicals of individual intelligent agents.

RBF network: This is a kind of feed forward network tool for training supervised modules. This is a kind of modification of back propagation network with a single hidden layer. Unlike sigmoid function like back propagation network, for activation it uses gaussian kind of functions which is selected form a clsas of activation functions called basis function.

III. PERFORMANCE ANALYSIS OF EXISTING SYSTEMS

We have run the above discussed algorithms in data mining tool tangara. We analyzed the data and implemented the techniques to get a rough overview of competitive performance among the existing models.

Table-1:

Methodology	Accuracy
Naive Bayes	73%
Decision tree	68%
K-nearest neighbour	70%
Support Vector Machine	80%
Clustering (Foggy k-means)	75%

IV. PROPOSED MODEL

We aim to design an end to end analytical model for prediction of heart disease where in the final diagnosis we use a total of 14 attributes to predict the risk probability of cardio-vascular disease in patient's body. We tend to make our approach quite robust and scalable. It should accomplish the targeted goal of data science research i.e. to solve real world scenarios (e.g. Estimating the pattern and trends of heart disease in a region of a country).

V. TOOLS AND FRAMEWORK

CNN: Convolutional neural network is a type of deep learning network with size and shape invariablity. When it comes to image classification or graph analysis, CNNs outperform many other classification models as because of their little preprocessing. They use convolution functions for the inputs to send the processed data to the next layer.

Crystal: Crystal is an object oriented language with a high level scripting workability introducing type inference algorithm for determining the variable type. Crystal is in developmental state, but is highly useful for computing neural networks with multiple layers. For predicting the risk of heart disease over the dataset training on 14 health attributes, we have designed a multiclass convolutional neural network in crystal. Tensorflow: Developed by Google Brain team, Tensorflow is a machine learning framework extraordinarily used for deep learning implementation where one can construct multilayer and multiclass neural networks with data flow graphs. It typically uses tensors as data structure. Tensors are multidimensional array alike static type structures. Tensorflow represents all kinds of data (e.g: scalar, vector, matrices) through tensors. Tensors varries in their particular data type and maintain states through the complete execution of the graphical model.

VI. DESIGN OF MODEL

The project so far contains 2 modules:

1. ECG report analysis and prediction of atrial fibrillation with convolutional neural network.
2. Predicting risk of heart disease by multiclass artificial neural network.

Module-1: ECG analysis:

Electrocardiography (ECG):

Electrocardiography depicts the electrical activity of a large mass of atrial and ventricular cells and to summarize, ECG actually records two types of events:

(1)Depolarization- the spread of electrical stimuly across the muscles of heart and (2)Repolarization: The event of return of stimulated muscles to the state of rest. ECG is expressed in a continuous graph with peaks and depressions and we will discuss about an structural unit of an ECG report below.

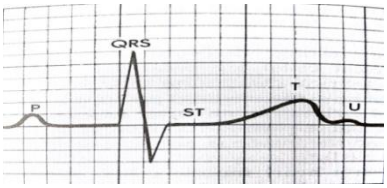


fig:1- model ECG wave segments[3]

ECG report:

Above is a model wave form structure of an ECG report. The graph comprises of 3 complex or specified waves which signifies 3 discrete events.

Table-2:

Complex	Clinical significance
P-wave	Atrial depolarization
QRS complex	Ventricular depolarization
ST segment, T wave, U wave	Ventricular repolarization

The data:

The dataset used in our project is MIT-BIH arrythmia dataset which is downloadable from Physionet official site [9].The data contains two channel ambulatory ECG recording of 48 patients. The detailed description of the dataset can be accessed through the source mentioned.Input requirement of our model is in .mat format. One should refer to the research paper describing the data [10].

Analytical model:

The model is an 1-D convolutional network, with a kernel size 16, and 64 filters and a default stride of 1.The neural network is built in python with Tensorflow framework. It takes real-time or consolidated ECG result fetched by any smart fitness device or ECG machines as input. We use the MIT-BIH chest arrythmia database provided by Physionet.org for training and testing the model . We split and merge the whole dataset into 2 parts one for training purpose and the other for testing the model. The operation is done randomly by a 'merge-dataset' program written in python. We run the train dataset first to train the model on the given files consisting of the ECG results and detection remarks and then try the model on test.mat file to check the accuracy, prediction result, false positive and false negative score, the recall, precision and hence calculate the F-1 score. After repeated run and 24 fold cross validation our model has reached an F1 score of .86 which is quite better when compared to the existing models. The model detects possible existance of atrial fibrilation in patients' heart, provided his ECG result as input. We have built 2D convolutional network with the input ECG leads transformed into a 2D array. The 2D convolutional network comprises of 2 layers, one with 32 and the other with 64 feature maps with 1 x 5 shape. Each layer has a max pooling index of 2. In our developed CNNs we have used ReLU as actiavtion function of neurons with categorical cross entropy loss. To fight the problem of overfitting we have applied batch normalization and with a dropout rate of 0.5. The 2D convolutional neural network,if applied on the transformed dataset, promises an F1 score of more than 0.92 as experimented in our project.

Module-2: Predicting risk of heart disease:

After being done with the proximal analysis of patients' ECG results, we move on to the next part of our project which is the prediction of risk of cardio vascular diseases for which we use a deep learning approach on a multiclass neural network which we train upon a typical dataset which consists of 14 health attributes.

The data:

The dataset is derived from UCI data repository. It has a consolidated research result of 303 patients across 3 countries and their respective four eminent hospitals. Each instance has 76 attributes out of which 14 are the principle ones and others are derived as subset of the former. Out of 14 attributes, 2 are holistic for prediction (i.e: Age, Sex), 11 are numerical and graded attributes and 1 is a predicted attribute, derived from 10 class attributes and signifies the angiographic disease status by the extent of diameter narrowing of vessels (less than or greater than 50%). For the detailed description of the dataset, one should refer to the UCI heart-disease dataset [11].

Predictive model:

We have built a multiclass deep neural network written in crystal . In stead of using matrices, it uses object oriented method in building a neural network. Our multilayer model has 10 input layer nodes, 10 hidden layer nodes, and 2 output layers following the convention :

number of hidden layer nodes = 2/3* number of input layer nodes + number of output layer nodes.

We provide the vivid description about the specifications of our model below.

Sigmoid activation function:

As an activation function, attached to the output layer of a neural network generally determines the type of the output the layer ejects, we have carefully chosen a sigmoid of S-shaped activation function for the implementation of our model because sigmoid activation functions are the best ones used for predictive output of multiclass neural networks. Our desired output is a possibility index of risk probability of heart disease which can vary from 0 as the lowest to 1 as highest considering all possible intermediate values as depicted in fig-2. Sigmoid functions are also differentiable and monotonic which adds an advantage for processing of our main function.

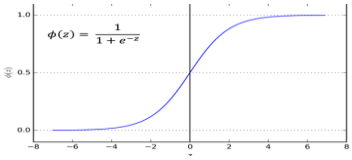


fig:2- Sigmoid activation function

Mean Square Error cost function:

The MSE cost function measures the mean square average of the outputs of a predictive model those vary from the correct output. The cost increases when the performance of the output is worse on the training dataset.

It is generally expressed by J(W) for a parameter W which minimizes the value of the whole function thus making the performance of the model better.

Taking W as weights, X as the input vector at ith training example and y as the class level of the ith training example and hw(xi) is the prediction using the ith training example we define J(W) as:

$$J(W) = 1/m(h_w(x)-y)^2 \text{ - summation of all positive values from 1 to m}$$

-where m signifies the number of training example (330 in our case).

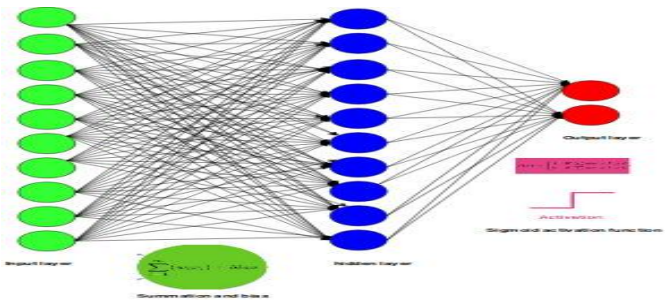
SGD with momentum optimization:

As our learning optimizer, we have chosen the SGDM or stochastic gradient decent with momentum for training our model because it has shown quite promising performance in deep learning rather than using classic SGD functions by accelerating the gradients towards right direction by navigating through the narrow ravines near the local minimas where classical SGD oscillates near the ravines and gets diverted towards the steep sides. The momentum actually represents the moving average of the gradients. Considering 'a' as learning rate (0.01 in our case), 'G' as gradient, 'W' as weight, 'L' as loss function and B as scaling factor we can define an SGD+momentum optimizer as:

$$V_t = Bv_t\text{-}1 + (1\text{-}B)G_W L(W,X,y)$$

$$W = W - aV_t$$

Architecture of the neural network:



Here below fig-3 is the schematic architecture of the neural network that we implemented.

fig-3: Architecture-neural network

VII. RESULTS

We have trained the model on the above mentioned dataset and have tested with 30 real time unseen patients' data. The model has a promised accuracy of 97% on the unseen patients' data. As defined we derive accuracy:

Accuracy = (tn+tp)/(tn+tp+fn+fp)

= (12+17)/(12+17+1+0) = 0.97 (our result)

-where tn is number of true negatives, tp is that of true positives and fn and fp are respectively the numbers of false negatives and false positives. The denominator equals to total number of output predictions.

VIII. CONCLUSION:

Through out the paper, we have evaluated and discussed about both existing and proposed methods, statistical models and systems for analysing ECG signals and predicting risk of heart disease evaluated on attributes. Further research will undergo on adding more analytical model of other diagnostic tests for detection of other types of heart related disease. The tools are developed for exclusively research purpose and not for diagnostic decision making. Active contribution towards the betterment of performance, reliability and expansion of the research is welcome. One can drop a mail on any of the authors' e-mail i'ds. For implementational details one should refer to our repository in GitHub (https://github.com/anshs99/Cardio-predict). We wish this model to be developed further and become a robust platform.

REFERENCES:

1. S. Vijayarani, S. Sudha, "Comparative Analysis of Classification Function Techniques for Heart Disease",International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 3, May 2013.
2. R.Ade, Dhanashree, S. Medhekar, Mayur P. Bote, "Prediction using SVM and Naïve bayes", International Journal of Engineering Sciences and Research Technology, May 2013.
3. "Clinical Electrocardiography- A simplified approach" by Ary L. Goldberger.
4. "Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks" by Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, Andrew Y. Ng.
5. Ashraf Osman Ibrahim, Siti Mariyam Shamsuddin, Nor Bahiah Ahmad, Sultan Noman Qasem "Three-Term Backpropagation Network Based On Elitist Multiobjective Genetic Algorithm For Medical Diseases Diagnosis Classification" Life Science Journal 2013. Pp 1815-1823.
6. T Hastie, R Tibshirani, and J H Friedman. The elements of statistical learning, volume 1. Springer New York, 2001.
7. M. Anbarasi, E. Anupriya and N.Iyengar, "Enhanced prediction of heart disease with feature subset selection using Genetic algorithm", International Journal of Engineering Science
8. and Technology vol.2, pp.5370- 5376, 2010.
9. Nidhi Bhatla, Kiran Jyoti, "An Analysis of Heart Disease Prediction using Different Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 8, October – 2012.
10. (https://www.physionet.org/physiobank/database/mitdb/) - Physiobank database source of MIT-BIH dataset.
11. Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. *IEEE Eng in Med and Biol* 20(3):45-50 (May-June 2001). (PMID: 11446209).
12. (https://archive.ics.uci.edu/ml/datasets/heart+Disease) – UCI machine learning repository- heart disease dataset.
13. "Machine Learning with Convolutional Neural Networks in Medical Diagnosis" by Michael Smith- 200784771, MPhys Research Project.
14. "Predicting Cardiac Disease With Deep Learning" by Taylor Archibald, Corey Woodfield, Jesse Robinson, and Benjamin Bay- December 2017 CS 478–Machine Learning Brigham Young University.
15. "Principal component analysis. Chemometrics and intelligent laboratory systems" by Svante Wold, Kim Esbensen, and Paul Geladi.2(1-3):37–52, 1987.

AUTHORS PROFILE

Soumonos Mukherjee, graduated from VIT,Vellore with a B.Tech in computer science and engineering with specialization in bioinformatics. Received post graduate certification on Scalable Data Science from IIT,Kharagpur. Cofounding a supply chain analytics startup Tekmeda solutions .

Anshul Sharma, pursuing B.Tech in Computer science and engineering from VIT,Vellore.